20. Logistic Regression

We continue to study the relationship between a response variable and one or more explanatory variables. For SLR and MLR (Chapters 18 and 19), our response was continuous or, at least, numerical (e.g., counts). An important assumption for linear regression is a normally distributed response variable (or error). Now we study a "regression" procedure appropriate for binary responses (e.g. Y=0 or Y=1): logistic regression. The explanatory variables can still be either categorical (e.g., indicator) or numerical variables. Clearly, the normal distribution is not appropriate for modeling such a random variable.

Recall, for regression, we modeled the mean of our response variable y as

$$\mu_{y|x_1,x_2,\ldots,x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

Also, recall that the mean of a binary (Bernoulli(p) or binomial(1,p)) random variable is just p, the probability of a "success"--P(Y=1).

As a first (naïve) attempt at modeling the mean, p, as a function of explanatory variables, we might write:

$$p_{x_1,x_2,\ldots,x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

But…

E.g. Aortic Stenosis, Smoking and Gender (Exercise 8, Chapter 20) How does aortic stenosis vary with smoking and

gender? Since disease status (aortic stenosis/no aortic stenosis), smoking (yes/no), and gender (male/female) are each categorical variables, we could use the contingency table methods of Chapter 15 and 16. In fact, the contingency table methods are closely related to the logistic regression method we'll discuss—contingency table methods are often used to introduce the logistic regression method.

Here, our response variable Y indicates whether a subject has aortic stenosis (Y=1) or not (Y=0); Y~bin(1,p):

P(subject has aortic stenosis—"disease")=
P("success")=
P(Y=1)=p
P(Y=0)=1-p

Some questions we might ask:

- How does the probability, p, of disease change with smoking status?
- Does the probability of disease change with gender?
- Does the relationship between aortic stenosis and smoking status depend on gender?

Again, we might use the Mantel-Haenszel method to study the relationship of disease with smoking (test for homogeneity, and then, if appropriate, test for association or calculate the overall log odds ratio, remember?) Here, we use logistic regression to perform a similar analysis.

Instead of modeling the probability, p, as a linear function of our response variables, we model the log odds as a linear function of the explanatory variables:

$$\log(p/(1-p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

(Again, log = natural log here.) Note that log(p/(1-p)) is called the logit of p, commonly written as logit(p). So we could write

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

Logistic regression is a particular case of a more general methodology called generalized linear models (GLM). SLR and MLR are also particular cases of GLM. We "generalize" to modeling some function of the mean—logit(p) for logistic regression—as a linear function involving our explanatory variables. Similar to SLR/MLR, we'll get an estimated equation (not least squares now):

$$\text{logit}(\hat{p}) = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_q x_q$$

which we can rearrange to get

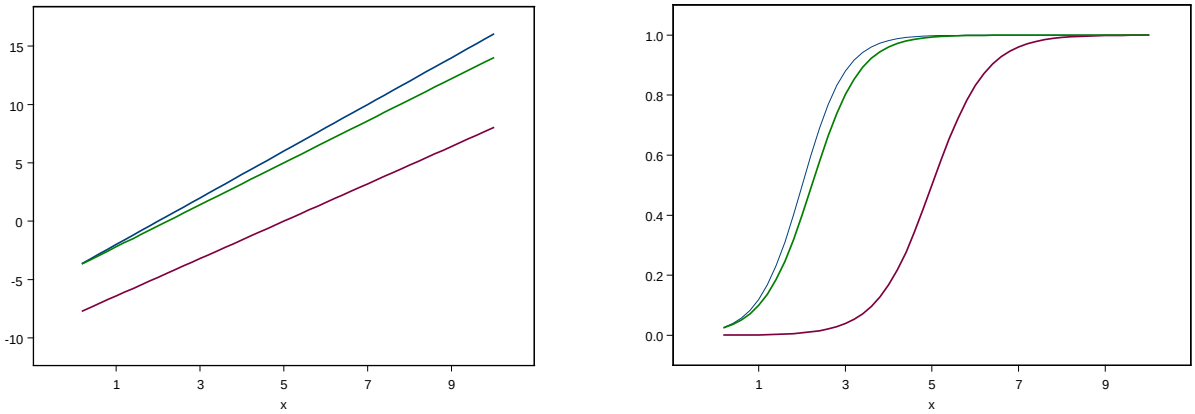
$$\hat{p} = \frac{e^{\hat{\alpha}+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\cdots+\hat{\beta}_q x_q}}{1+e^{\hat{\alpha}+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\cdots+\hat{\beta}_q x_q}}$$

Before we give the analysis, let's inspect the logistic regression model further. We'll use one (continuous) explanatory variable.

$$\text{logit}(p) = \alpha + \beta x$$

or

$$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$



- β is still the slope parameter, but is interpreted in the context of log(odds)—increase the explanatory variable by 1 unit, and β is the (additive) change in the log(odds). That is, the log(odds) of disease for a subject in the "group" having explanatory variable value x+1 is changed by β compared to a subject from the "group" having explanatory variable value x
- Often, we prefer to talk of odds rather than log(odds), so we exponentiate. In this case we have the following change in the odds from x to x+1

$$e^{[(\alpha+\beta(x+1))-(\alpha+\beta x)]} = e^{[\beta]}$$

where the change in the odds is multiplicative. To see this, we look at the above as

$$e^{[\log(\text{odds}_{x+1})-\log(\text{odds}_x)]} = e^{[\log(\frac{\text{odds}_{x+1}}{\text{odds}_x})]} = \frac{\text{odds}_{x+1}}{\text{odds}_x} = \text{oddsratio},$$

which is $e^{[\beta]}$. Thus, $e^{[\beta]}$ is the odds ratio, or the multiplicative change in odds from the x "group" to the x + 1 "group".

E.g., continued

Our example is a special case of logistic regression with indicator explanatory variables (they're categorical). Again, we could have used contingency table methods.

n=215 subjects (from stenosis.xls; see also section 16.2)

Smoke no=0 yes=1	disease no=0 yes=1	gender female=0 male=1
1	1	1
1	1	1
1	1	1
…snip…		
0	0	0
0	0	0
0	0	0

Statistics>Regression>Logistic…

```
       *** Generalized Linear Model ***

Call: glm(formula = disease ~ smoke, family = binomial(link = logit), data = stenosis,
na.action = na.exclude, control
        = list(epsilon = 0.0001, maxit = 50, trace = F))
Deviance Residuals:
       Min        1Q    Median       3Q      Max
 -1.250676 -1.087288 -1.087288 1.188073 1.270281

Coefficients:
                Value Std. Error   t value
(Intercept) -0.2157085  0.1828685 -1.179582
      smoke  0.3863340  0.2762301  1.398595

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 297.937 on 214 degrees of freedom

Residual Deviance: 295.9722 on 213 degrees of freedom

Number of Fisher Scoring Iterations: 2
```

- log(odds) of disease increases (additively) by $\hat{\beta}$=0.386334 for those who smoke verses those who don't smoke, or

- odd of disease increase (multiplicatively) by exp(0.386344)=1.4716 times for those who smoke verses those who don't smoke, or
- in other words, the estimated odds ratio is 1.4716
- compare to results on page 375—this is not a coincidence

But, our explanatory variable (smoking indicator) does not seem to be significant. Also, we should be thinking about how the relationship between disease and smoking is affected by gender, as we were in Chapter 16…more shortly…first some testing.

As with SLR and MLR, we can test hypotheses about coefficients in the linear predictor using a normal distribution approximation:

Ho:
Ha:

test statistic and approximate p-value

An alternative test procedure is the "drop in deviance" approach, which may be a better approximation for testing. We merely illustrate this approach but do not discuss what deviance is except to say that it is analogous to residual sum of squares in the linear regression case:

How does gender affect our conclusions about disease and smoking? Now we take a Mantel-Haenszel sort of approach to our example data (see Chapter 16).

First, test for homogeneity of the effect of smoking on disease across gender (test for interaction of smoke and gender). That is, does the change in log(odds) of the disease for smokers vs. non-smokers depend on gender?:

```
       *** Generalized Linear Model ***

Call: glm(formula = disease ~ smoke + sex + smoke:sex, family = binomial(link = logit), data
= stenosis, na.action =
       na.exclude, control = list(epsilon = 0.0001, maxit = 50, trace = F))
Deviance Residuals:
       Min        1Q     Median       3Q      Max
 -1.365882 -1.050779 -0.9803974 1.084239 1.388115

Coefficients:
                Value Std. Error     t value
(Intercept) -0.48284195  0.2358803 -2.04697858
      smoke  0.17746071  0.4238581  0.41867959
        sex  0.70598545  0.3816154  1.84999195
  smoke:sex  0.03225522  0.5815873  0.05546067

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 297.937 on 214 degrees of freedom

Residual Deviance: 289.6403 on 211 degrees of freedom

Number of Fisher Scoring Iterations: 2
```

We don't reject the null of homogeneity across gender. In the M-H procedure of Chapter 16, we would continue with estimating the overall (log) odds ratio ("combining" information across levels of gender) or conducting a test for association between disease and smoking ("combined" across gender):

Similarly, in the logistic regression setting, we can test the null of combined odds ratio = 1 (after adjusting for sex), which is equivalent to testing:

Ho:
Ha:

```
       *** Generalized Linear Model ***

Call: glm(formula = disease ~ smoke + sex, family = binomial(link = logit), data = stenosis,
na.action = na.exclude,
       control = list(epsilon = 0.0001, maxit = 50, trace = F))
Deviance Residuals:
       Min         1Q     Median       3Q      Max
 -1.362978 -1.055554 -0.9783289 1.080724 1.390484

Coefficients:
                Value Std. Error   t value
(Intercept) -0.4881595  0.2157069 -2.263069
      smoke  0.1945917  0.2901686  0.670616
        sex  0.7198847  0.2879623  2.499927

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 297.937 on 214 degrees of freedom

Residual Deviance: 289.6434 on 212 degrees of freedom

Number of Fisher Scoring Iterations: 2
```

We don't reject the null of no association between disease and smoking (after accounting for gender). We try a final, simple model using only gender as a predictor:

```
       *** Generalized Linear Model ***

Call: glm(formula = disease ~ sex, family = binomial(link = logit), data = stenosis,
na.action = na.exclude, control
        = list(epsilon = 0.0001, maxit = 50, trace = F))
Deviance Residuals:
       Min        1Q    Median       3Q      Max
 -1.326084 -1.001693 -1.001693 1.035669 1.363925

Coefficients:
                 Value Std. Error   t value
(Intercept) -0.4284503  0.1958458 -2.187692
        sex  0.7713941  0.2778310  2.776487

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 297.937 on 214 degrees of freedom

Residual Deviance: 290.0919 on 213 degrees of freedom

Number of Fisher Scoring Iterations: 2
```

What's the estimated log odds of disease for men?

For women?

The estimated odds ?

The estimated probability of disease?

What's the difference in estimated log odds between men and women?

What's the estimated odds ratio of disease of men verses women?

Give an approximate 95% CI for the difference in log(odds) of disease for men vs. women.

Give an approximate 95% CI for the odds ratio for men vs. women.

E.g. Peritonitis, age, sex, and race (Exercise 9, Chapter 20, page 487.